AFRL-RH-WP-TR-2011-0060

CULTURE & COGNITION LABORATORY

Rik Warren

Anticipate & Influence Behavior Division
Behavior Modeling Branch

MAY 2011
Final Report

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
HUMAN EFFECTIVENESS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2011-0060 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

RICHARD WARREN
Work Unit Manager
Behavior Modeling Branch

DAVID G. HAGSTROM
Anticipate & Influence Behavior Division
Human Effectiveness Directorate
711th Human Performance Wing
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* 02-05-2011	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* June 2005 – April 2010

4. TITLE AND SUBTITLE Culture & Cognition Laboratory	**5a. CONTRACT NUMBER** In-House
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER 62202F
6. AUTHOR(S) Rik Warren	**5d. PROJECT NUMBER** 7184
	5e. TASK NUMBER 10
	5f. WORK UNIT NUMBER 71841009
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Anticipate & Influence Behavior Division Behavior Modeling Branch	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Materiel Command Air Force Research Laboratory 711th Human Performance Wing Human Effectiveness Directorate Anticipate & Influence Behavior Division Behavior Modeling Branch Wright-Patterson AFB OH 45433-7022	**10. SPONSOR/MONITOR'S ACRONYM(S)** 711 HPW/RHXB **11. SPONSOR/MONITOR'S REPORT NUMBER(S)** AFRL-RH-WP-TR-2011-0060

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

88ABW/PA cleared on 24 May 2011, 88ABW-2011-2868.

14. ABSTRACT

This is the final report for research done under in-house Workunit 71841009 entitled the Culture & Cognition Laboratory (CCL) during the period 5 June 2005 through 7 April 2010. Although CCL is an Air Force Research Laboratory (AFRL) Human Effectiveness Directorate research facility, it is located off-site at Wright State University (WSU) per a Cooperative Agreement for laboratory space and a Cooperative Research and Development Agreement for general research support. The report includes descriptions of the physical aspects of CCL, its capabilities, the Situation Authorable Research Environment (SABRE) used to conduct experiments, descriptions of experiments using a computer role-play game which enables complex problem solving by actively engaged interacting players from different cultures, a review of behavior modeling efforts under the workunit, and a novel statistical techniques measuring maximal diversity in teams. The report concludes with an extensive set of lessons learned and recommendations.

15. SUBJECT TERMS

Culture, Cognition, Behavior Modeling, Cooperation, Serious Games, Agent-Based Simulation

16. SECURITY CLASSIFICATION OF: UNCLASSIFED			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Richard Warren
c. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	SAR	42	**19b. TELEPHONE NUMBER** *(include area code)* NA

THIS PAGE LEFT INTENTIONALLY BLANK

TABLE OF CONTENTS

List of Figures

THIS PAGE LEFT INTENTIONALLY BLANK

SUMMARY

This is the final report for research done under in-house Workunit 71841009 entitled the *Culture & Cognition Laboratory* (CCL) during the period 5 June 2005 through 7 April 2010. Although CCL is an Air Force Research Laboratory (AFRL) Human Effectiveness Directorate research facility, it is located off-site at Wright State University (WSU) per a Cooperative Agreement for laboratory space and a Cooperative Research and Development Agreement for general research support. The report includes descriptions of the physical aspects of CCL, its capabilities, the Situation Authorable Research Environment (SABRE) used to conduct experiments, descriptions of experiments using a computer role-play game which enables complex problem solving by actively engaged interacting players from different cultures, a review of behavior modeling efforts under the workunit, and a novel statistical techniques measuring maximal diversity in teams. The report concludes with an extensive set of lessons learned and recommendations.

1 INTRODUCTION

This is the final report for research done under in-house Workunit 71841009 entitled the *Culture & Cognition Laboratory* (CCL) during the period 5 June 2005 through 7 April 2010. Although CCL is an Air Force Research Laboratory (AFRL) Human Effectiveness Directorate research facility, it is located off-site at Wright State University (WSU) per a Cooperative Research and Development Agreement. Integral to the research program were two Cooperative Agreements with WSU for research and technical support to enable data collection and analysis. Accordingly, this report documents matters of the Cooperative Agreements as well as the conduct and results of the CCL research program.

1.1 Culture & Cognition Laboratory: Purpose

The purpose of the Culture & Cognition Laboratory was (and is) to study the impact of cultural factors on cognition, decision-making, and collaboration.

Fulfilling this purpose entailed the establishment and maintenance of a CCL facility as well as the development and execution of a coherent research program.

1.2 Culture & Cognition Laboratory: Facility Background

After the attack on the United States on 11 September 2001, and subsequent developments in Iraq and Afghanistan, it became all too apparent that we needed to better understand the influence of culture on the thinking processes of allies and adversaries and its role on conflict and cooperation. Accordingly, it was decided to establish an in-house house laboratory for research on culture and cognition.

1.2.1 Need for CCL as an Off-Site Facility

Since research on cultural factors requires the use of large numbers of diverse foreign nationals as research subjects, and since access to an on-base research facility by foreign nationals in the large quantities needed for relatively short durations was (and is) essentially impossible, it was soon realized that an off-site research facility with easy access to large numbers of foreign nationals was needed. AFRL researchers thus decided to locate CCL "outside the gates." Establishing such a facility required entering the two agreements discussed below.

1.2.2 Location & Establishment of the CCL: CRDA with WSU

Since Wright State University is located within a short drive from the Air Force Research Laboratory at Wright-Patterson Air Force Base and since WSU has large numbers of foreign nationals in its student body, the CCL was established at WSU under a Cooperative Research and Development Agreement (CRDA No. 05-096-HE-04). The CRDA commenced on 6 April 2005 with an original expiration of 6 April 2010. The CRDA has since been amended with a new expiration date of 6 April 2015. A key feature of the CRDA is the provision for laboratory space on the WSU campus. The facility itself is described in a following section of this report.

1.2.3 Research & Technical Support for the CCL

An off-site research facility requires research and technical support especially since it is removed from the venue of the standard on-site support contracts.

Since the Wright State University Department of Psychology had (and continues to have) faculty expertise in cultural psychology and cognitive psychology, graduate students trained in research on individual differences and teamwork within its Ph.D. program in Industrial/Organizational and Human Factors Psychology, and a skilled computer programming staff, it had the necessary resources to provide a wide range and depth of research and technical support to the CCL. Thus, a Cooperative Agreement (FA8650–05–2–6625) entitled *Research and Technical Support for the Culture & Cognition Laboratory* was awarded to WSU effective June 3, 2005 with an original duration of 12 months. Subsequent amendments extended the expiration of the technical effort to April 7, 2010. A second Cooperative Agreement (FA8650–10–2–6132) entitled *Research and Analytic Support for the Culture & Cognition Laboratory* was awarded to WSU effective April 13, 2010 with an original duration of 15 months.

1.3 Structure & Scope of the Report

Because of the integral relationship of the research facility that is the Culture & Cognition Laboratory, and the principal methodology for data collection, this report documents these prior to descriptions of the research studies and the lessons learned as a consequence of the research program. This report includes the following discussions:

- The research facility. This is a description of the physical aspects of the CCL.
- Research capabilities and philosophy.
- In order to understand the major data collection efforts, a necessary first step is a familiarity with the main research methodology featured in CCL. The Situation Authorable Behavior Research Environment (SABRE) is a powerful tool for conducting experiments in which multiple persons interact using a role-play computer game.
- A description of a five-nation large NATO experiment with military subjects and using CCL and SABRE.
- A description of an experiment involving culturally-mixed dyads in a task requiring good communication and cooperation for completion.
- A review of behavior modeling efforts to account for the data including some counter-intuitive findings.
- A novel statistical technique for measuring maximal diversity in teams.

Lastly, the report concludes with a set of lessons learned and a set of recommendations.

2 CULTURE & COGNITION LABORATORY

Physically, the Culture & Cognition Laboratory (CCL) is located in a two-room suite in Fawcett Hall at Wright State University (WSU). See Figure 1.

The smaller of the two rooms contains the operator workstation/server for the CCL intranet. It also, houses two data-collection workstations.

The larger of the two rooms houses four workstations. Two workstations face a wall and two others face the opposite wall. Additionally, portable telescoping partitions separate the side-by-side workstations. The arrangement is such that no subjects can see each other unless they face the center of the room or peer around a partition. Each workstation is equipped with a 21" monitor, a mouse and keyboard to interact with the display, and a headset with microphone so subjects can communicate with each other but not be bothered by other possible noises in the room.

A network of seven computers serves as the operations center for CCL. One computer is configured as the operator workstation and server for the other six computers which are configured as subject workstations. The CCL network is interfaced to the Internet via the WSU wireless net. This Internet interface permits team experiments and data collection in situations in which the team members are physically located in several different countries yet function as a team as if they were all in the same room.

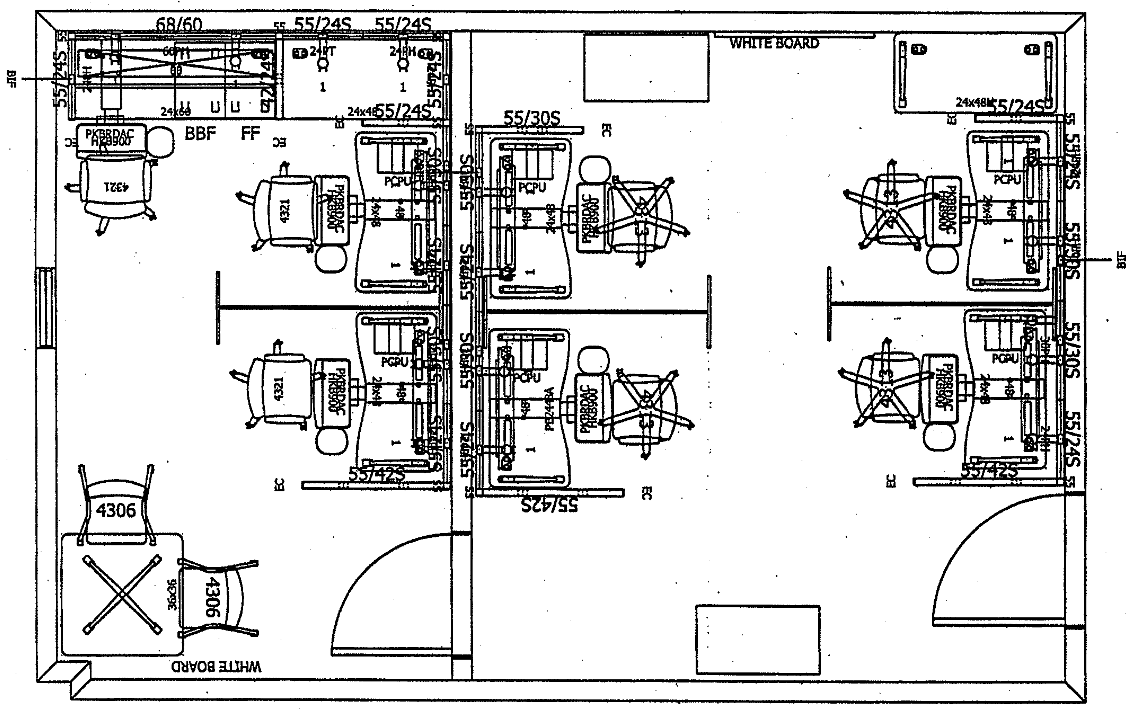

Figure 1: CCL Floorplan

3 RESEARCH CAPABILITIES & TOOLS OF CCL

Between CCL the facility and CCL the research program, there lies a set of software tools which enable a particular style and range of research studies. Those tools must be developed with the peculiar challenges and constrains of cultural research in mind. Ideally, a new laboratory can use some traditional methods while implementing new methods which avoid problems with the old and permit new types of studies. But a laboratory should also have a philosophy of research that gives it a coherence and clear direction.

3.1 Challenges of Cross- & Inter-Cultural Research

Research on culture and cognition is notoriously difficult. Inter-cultural studies necessarily face the problem of comparing apples to oranges since different cultures often can and do view and react to the "same" circumstance differently. For example,

- Different cultures often use a different language or dialect. The meaning of words and phrases often do not easily translate with all the nuances and subtleties captured. This is the well-known problem of something being "lost in translation."
- Even when there is a reasonable semantic correspondence between research materials (such as instructions, scenario descriptions, and rating scale descriptions) used with two different language or cultural groups, differences in cultural values and norms can affect the way people respond. This is especially problematic with the very common data-gathering "pencil and paper" techniques such as rating scales. For example, in some cultures it is important to be seen as truthful and not hedging in giving one's opinion. Hence, such respondents tend to use the extreme ends of ratings scales. In other cultures, it is assumed that life is complex, that many things are inter-related, and that it would be immodest for someone to claim full knowledge of an area. Respondents from these cultures tend to favor the middle points of ratings scales. Although various statistical techniques have been proposed to remove the effects of cross-cultural response bias, they are not fully successful (Fischer, 2004). Another problem is that rating scales and similar techniques are vulnerable to deliberate image-manipulation by the responders.

Given the challenges, the next step was to review the methods used in cross-cultural and inter-cultural research.

3.2 Methods of Cross- & Inter-Cultural Research

An overwhelming number of cross-cultural and inter-cultural studies use passive pencil-and-paper techniques such as 7-point rating scales since they are easy to administer and score (Brouwers, Van Hemert, Breugelman, & Van de Viyer, 2004). But it is those very passive paper-and-pencil techniques which render the studies vulnerable to response biases and image manipulations.

One way to avoid the bias problems is to use active techniques in which members of different cultures actually interact with each other. However, such studies are logistically cumbersome, hard to conduct, expensive to record, and onerous to score to enable quantitative analysis. As a result, very few studies use active techniques (Brouwers et al., 2004).

3.3 Determining a Laboratory Active Research Methodology

It was clear that the best laboratory methodology needed to involve active interaction by the subjects, but that a way had to be found to simplify the associated logistical and procedural burdens, The next task, then, was to bound the problem space and develop a suitable laboratory methodology that would permit quantitative assessment of cultural effects on cognition while at the same time avoiding or vastly mitigating cross-cultural response bias.

The solution was to capitalize on an AFRL/RH-managed Defense Modeling and Simulation Office (DMSO)-funded contract with BBN Technologies to develop a "Game-based testbed for culture and personality research" (Leung, Diller, & Ferguson, 2005; Warren, Sutton, Diller, Ferguson, & Leung, 2004).

The genius of the DMSO-funded AFRL/RH-managed BBN-developed testbed was to use a serious role-play computer game to actively involve players on challenging tasks requiring interaction and communication. By using an absorbing game and task, the supposition is that:

- Players lose themselves in the game and let their true selves emerge. Further,
- Game scenarios can be developed which mask the hypotheses being tested by requiring players to make choices and permit taking a variety of actions. For example, if a player discovers a valuable item, does the player inform the other players?

Since all activity takes place in the "game space" and all communications and computer keystrokes can easily be recorded, the logistics problems are greatly reduced.

The particular tailoring and evolution of the game-based testbed for CCL use was termed the *Situation Authorable Research Environment* (SABRE).

4 SITUATION AUTHORABLE RESEARCH ENVIRONMENT

Because of the central role played by SABRE in influencing and enabling the CCL experiments, it is briefly described here. Fuller descriptions are in Leung, Diller, and Ferguson (2005) and Warren, Diller, Leung, Ferguson, and Sutton (2005).

4.1 Game Engine: *Neverwinter Nights*^{TM}

The heart of SABRE is the multi-player role-play computer game *Neverwinter Nights*^{TM} (Bioware, 2004). *Neverwinter Nights*^{TM} is popular among researchers since it has an Application Programming Interface (API) and powerful scripting language that permit easy tailoring, scenario generation, and access to almost all internal state variables for data-collection. The game can be "re-skinned" to permit scenarios in various time-periods and environments (e.g., forests, towns, the insides of buildings). An active user community exists who contribute many elements for others to use. The variety of capabilites and scenarios is essential unlimited and just depends on the ingenuity of the programmers.

4.2 Game Play Aspects

Each human plays at their own individual computer workstation, which includes a keyboard, mouse, and microphone headset. Each player has an avatar (a personal representative or alter-ego) in the game-space. The various avatars (and thereby their individual human controllers) can interact and communicate with each other. If avatars are in the same area, they can "talk" directly in the game-space, but if they are far removed, they can communicate by "radio." The avatars can pick up, carry, and otherwise interact with certain inanimate objects such as maps, tools and weapons. They can perform actions such as walking, running, opening doors, or operating control levers. In addition to the avatars, there can be various non-player (i.e., computer generated) characters (NPC's) which also live in the game-space and which can be interacted with and communicated with. The NPC's can be given various cultural traits and personalities. For example, they can lie in response to questions. The NPC's can also initiate conversations and ask questions of the avatars.

Since the scenarios, tasks, and manipulation skills (picking up objects, opening doors) can be complex, a within game training phase is often included in which a human, via an avatar, works with one or more NPC trainers who teach and drill the skills until some level of proficiency is reached.

4.3 SABRE as a Research Environment

As its name implies, SABRE is a general purpose research environment, not just a particular skin and scenario overlay on a computer game. This means that SABRE has facilities for data gathering, aggregation, and output to databases in formats suitable for later data analysis.

Not all data collection occurs within the game-play. For example, an experimenter often wants to obtain demographic information and administer pre- and post-game questionnaires and scales which are presumed to depend on a subject's culture. These can include, for example, a scale to measure analytic versus holistic thinking (Choi, Koo, & Choi, 2007) or a measure of positive and negative affect (Watson, Clark, & Tellegen, 1988). SABRE can integrate these for computer presentation, scoring, and database output. This procedure has several advantages such as no missing data (since the computer can flag when an item is omitted and require an answer to proceed) and no scoring errors.

In addition to the out-of-game questionnaires, and scorable within-game activities (e.g., number and types of items found, number and types of communications with other players and NPC's), in-game probes can be subtly or blatantly inserted. For example, an NPC can approach an avatar and ask an innocent question (e.g., "Excuse me, how do I get to the gas station?") that assesses the player's situation awareness. Every keystroke is logged for subsequent analysis.

The "Situation Authorable" aspect of SABRE is also very general as can be seen in two very different CCL experiments which used SABRE for scenario and task development.

4.4 Principal Experimentation Scenarios

In its basic form, Neverwinter Nights is a medieval fantasy, but for CCL use, various modern "skins" and scenarios were developed.

The first SABRE scenario revolved around a search-for-contraband involving a team of four players (Warren, Diller, Leung, Ferguson, & Sutton, 2005). The search took place in a modern urban setting where weapons caches were hidden both outdoors and inside buildings. Team members had to find as many weapons caches as possible without angering the local populace. This four-person scenario was used in the large NATO experiment described in the following section of this report.

Since the NATO scenario was very resource intensive – it required all day for a team of four to play – a second, shorter, scenario was developed for another series of experiments. The second series was designed to test the effects of the cultural composition of teams of two (dyads) on their evolving progress in solving a series of puzzles over a two-hour span. The puzzles could not be solved by a single person. Rather, they required the members to cooperate and exchange information in order to succeed. This scenario is also described later in this report.

5 PRINCIPAL EXPERIMENTS

Although the CCL facility is designed to be multipurpose and capable of hosting a variety of experiments, the research program is not an inchoate collection of studies whose only common denominator is the data-gathering location. The research on culture and cognition has followed a classical scientific cycle of hypothesis, test, evaluation, and new hypothesis. As can be expected in the relatively new area of culture and cognition, there has been partial support for some hypotheses, but more productively, several enriching surprises.

Two principal experimental scenarios have emerged from the progressive decisions, hypotheses, and new ideas during the period of performance. The resulting experiments have provided the data and inspiration for the modeling efforts discussed later.

5.1 NATO Experiment on Culture and Team Adaptability

In order to investigate the performance of mixed- versus homogeneous-culture military teams, the NATO Research and Technology Organization Research task Group HFM–138/RTG on "Adaptability in Multinational Coalitions" conducted a multinational experiment centered at CCL (NATO RTO HFM–138/RTG, 2008). The experiment itself used a SABRE-developed scenario and task (Warren, Diller, Leung, Ferguson, & Sutton, 2005).

The rational behind the experiment was that good communication is crucial for (possibly geographically distributed) team members conducting a complex military task such as searching for hidden weapons in an urban environment. It seemed reasonable to presume that effective communication should be best for people who share a common culture. Hence, the principal hypothesis was that teams whose members are all from the same nation perform better than teams whose members are from different nations.

In order to test the hypothesis, BBN Technologies used SABRE to develop a complex, but very absorbing and immersive urban search-for-contraband scenario. Successful performance required planning, resource allocation, situation awareness, good communication, and coordination. Good performance also required maintaining the good-will of the local "populace," that is, computer-generated characters who could provide useful or misleading information to the search team. Participants were 224 NATO officers formed into 56 four-person teams. Within a team, the four members were either from the same nation or were from four different nations. The game-play was all in English and all communications was by keyboard.

The results were surprising in that, contrary to hypothesis, the mixed-culture teams generally performed better (NATO RTO HFM–138/RTG, 2008). Figure 2 shows the overall performance score of each of the 56 teams arranged by nation. It also shows boxplots superposed over each national grouping. The dark horizontal bars in the boxplots show that the mixed-nation teams had the best median score on the overall measure of performance. Figure 2 also shows that poorest performance was by the eight teams comprised of senior Norwegian officers. That their performance was poorer than the eight teams comprised of much younger Norwegian officers offers a clue to explaining the results. The younger and older Norwegians come from the same culture, but clearly differ on age. This raised the question of confounds and covariates affecting the results.

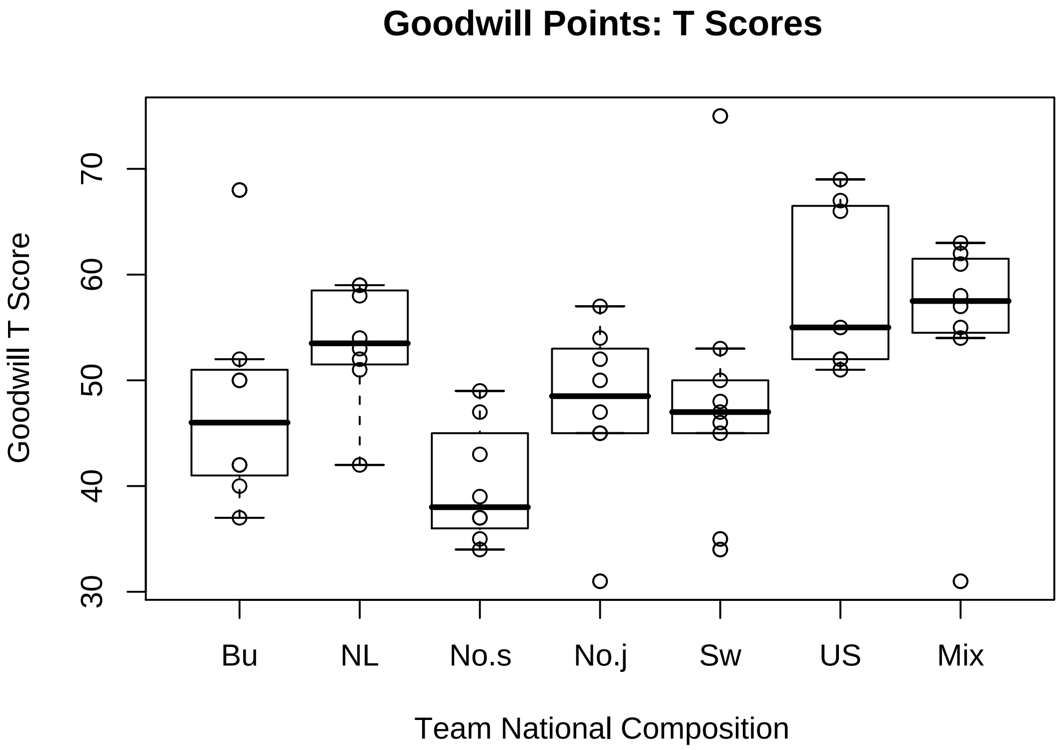

Figure 2: Overall game-play performance T-score (i.e., Mean = 50, SD = 10) for each of 56 teams grouped by national composition. Key: Bulgaria (Bu), The Netherlands (NL), Norway-senior age (No.s), Norway-junior age (No.j), Sweden (Sw), & the United States (US), Mixed culture (Mix).

In addition to age, two other variables which can affect game-play performance independently of culture are computer-game experience and English proficiency. Based on several questions in the demographics questionnaire, Warren (2008) devised metrics for English proficiency and computer-gaming experience. Figure 3 shows a bubble-plot/scatter plot of these three variables which can affect game-play performance. In the scatterplot, gaming experience is shown by the size of the bubbles.

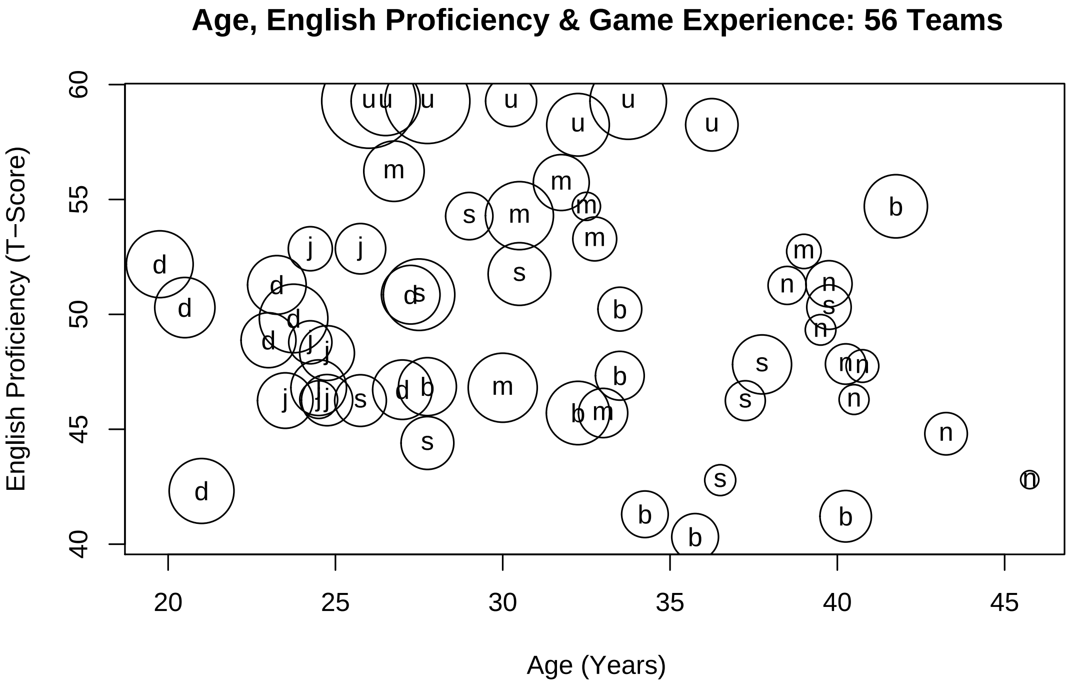

Figure 3: Demographic profiles of the 56 teams: Age, English proficiency, and computer-game experience. Game experience is proportional to size of bubbles. Letters indicate national composition of the teams. Key: Bulgaria (b), The Netherlands (d), Norway-senior age (n), Norway-junior age (j), Sweden (s), & the United States (u), Mixed culture (m).

It should be noted that the numerical value of all three covariates (age, English proficiency, and gaming experience) were all unknown prior to a subject reporting to the experiment and answering the questionnaires. That is, there was no way to assign subjects to teams or groups prior to scheduling all the members of a team. The best that can be done is to perform a post hoc statistical adjustment of the data to partial out the effects of the unavoidable covariates.

Since the number of teams per group is small, Analysis of Covariance (ANCOVA) is inappropriate, so Warren (2008) used multiple-regression to partial out the covariate effects. After the effects of the covariates (or confounds) of English proficiency, age, and computer-game experience were removed, the superiority of the mixed-culture teams was even more pronounced (See Figure 4).

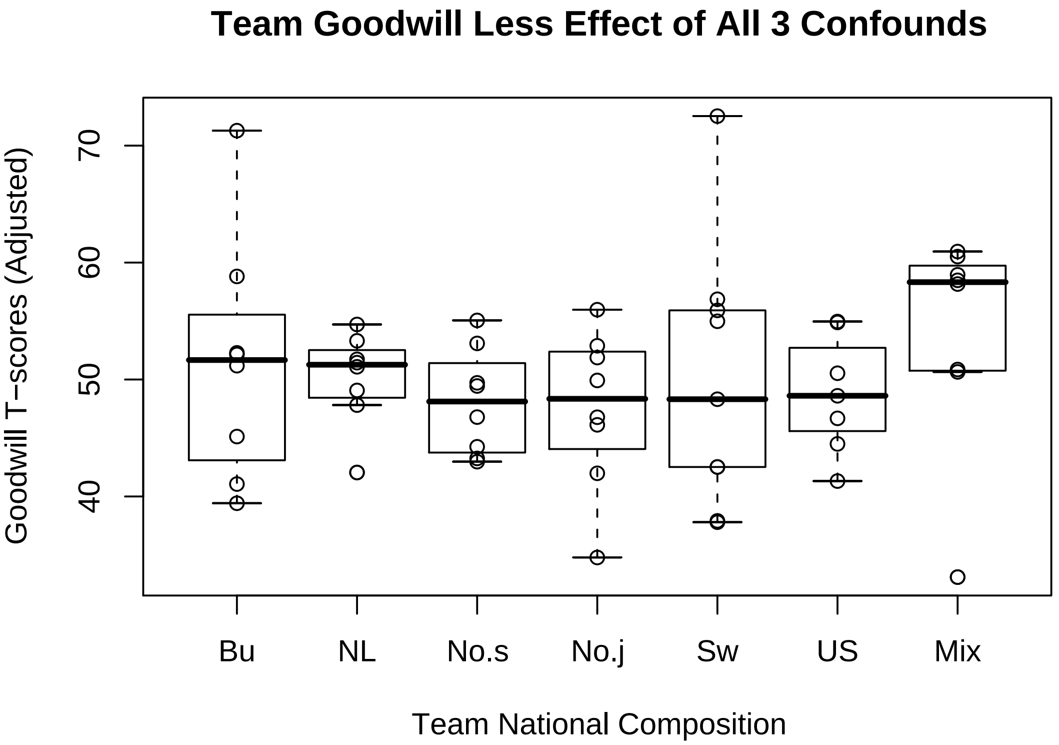

Figure 4: Game-play performance less effects of all 3 confounds (Adjusted T-scores, i.e., Mean = 50, SD = 7.87) for each of 56 teams grouped by national composition. Key: Bulgaria (Bu), The Netherlands (NL), Norway-senior age (No.s), Norway-junior age (No.j), Sweden (Sw), & the United States (US), Mixed culture (Mix). Compare with Figure 3. Box plots superposed on culture groups.

This counter-intuitive finding motivated a search for an alternative explanation based on the notion of the "wisdom of crowds" (Surowiecki, 2004/2005). The resulting analysis of the mechanism by which cultural diversity can have both positive and negative effects led to a two-factor opposing-mechanism model of diversity effects during complex tasks. This model was tested using an agent-based simulation and found to account for the counterintuitive effects (Warren, 2010). The new model is treated in more detail in a following section in this report.

Several lessons learned resulted from the NATO experiment. These also are presented in a following section of this report and in Warren and Sutton (2008). Some of the lessons concerned mitigating logistical problems and refinements to the experimental design. These issues and the need to further test hypotheses about communication and inter-cultural co-operation versus conflict led to a completely new scenario and task but still using SABRE. The focus shifted from using teams of four to teams of just two people.

5.2 Dyadic Problem-Solving Experiment

Disaster relief teams may be formed quickly with previously unacquainted members coming from diverse cultural backgrounds. The effectiveness of such newly-formed mixed-culture teams in solving complex problems depends on how well its members communicate, cooperate, and share information. Initially such groups have no history. They might start with either high expectations of each other or with low expectations based on bad stereotypes. But whatever their initial conditions, as the group interacts there will be an evolving history and an accumulation of effects. Figure 5 shows the predicted progress in the relative quickness of a dyad in sequentially solving a series of puzzles which are ordered left to right as time progresses.

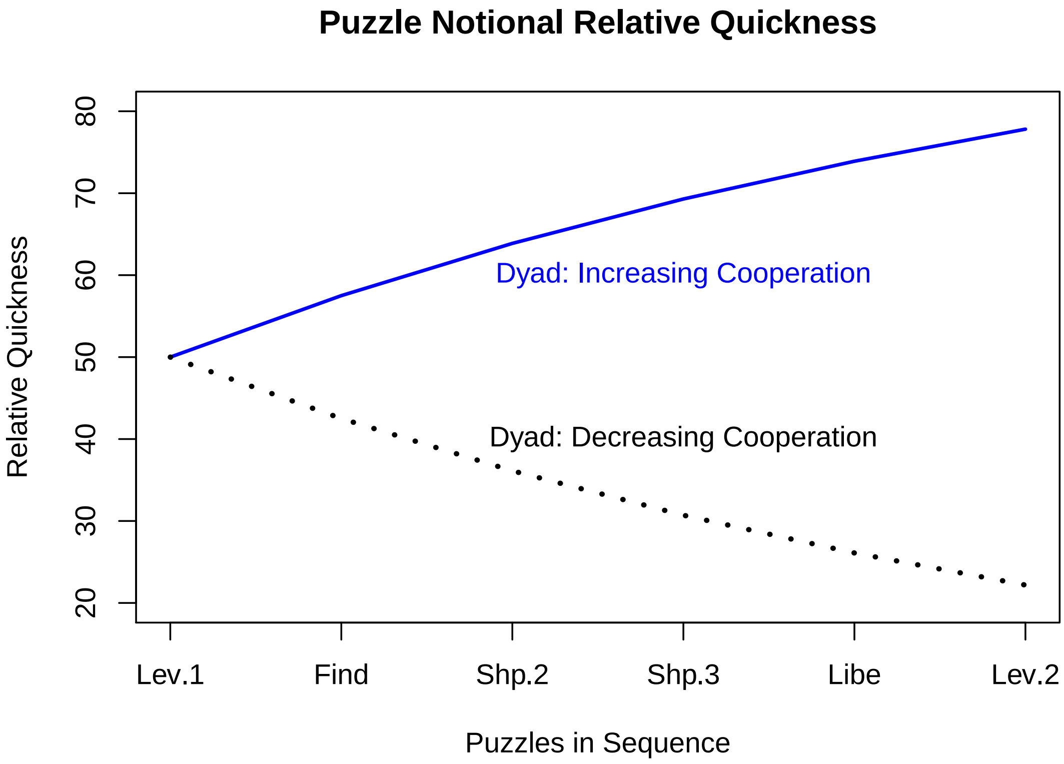

Figure 5: Predicted puzzle-solving effectiveness of an increasingly more cooperative team (upper curve) and an increasingly uncooperative team.

The dyad study was designed to assess the effects of group cultural composition on evolving problem-solving effectiveness. Dyads were chosen since that is the minimum group size possible. This helped mitigate a problem with the NATO study in which too often four subjects were needed, six were scheduled, and three showed up. The NATO study was complex and required almost a full day to run one team. The dyad study was designed to be challenging but require less training and total time to complete.

The solution was to have the dyads solve a series of puzzles which would be easy for a team to solve given good communication, but which would be impossible for one person to complete. The puzzles involved tasks for the players, via their avatars, to complete. For example,

- In four puzzles, players had to balance a ship by moving weights. In the most difficult version, one person was in a small locked room and could view a balance indicator. The other player could move freely about the ship. Since neither player could see each other, they had to devise ways to communicate and coordinate their activities. Consider that since the player in the indicator room could not see the other player, an instruction such as "move a weight a little to the left," by itself, makes no sense.

- In one puzzle, both players were on a mutual scavenger hunt, but only one player has the list of items to be gotten. Each player is charged with maximizing their own points, but at the same time, neither could carry more than a few items due to weight restrictions. The analogy is with sports: all must work together to win, yet one player wants to stand out to be the most valuable player. Hence, there is competition overlaid on cooperation.

- In two other puzzles, the avatars had to traverse a maze of rooms to reach a goal. The rooms and hallways had multiple doors which were locked such that a player could not pass unless the other player pulled a lever for them. This task required the players to take turns helping each other.

- In another puzzle, the players must exit a mansion before it burns down. One player is locked in a room with books containing information about where things, such as keys, can be found and how to get past obstacles. The other player is free to roam about the mansion. Good communication is required to successfully complete the task.

It was expected that homogeneous-culture dyads would communicate well and work smoothly, whereas it was possible that in some heterogeneous-culture dyads players could become frustrated with each other with cumulative negative effects on performance. To permit an assessment of the effects of positive or negative experiences, as part of the pre- and post-test questionnaires, players took the Positive and Negative Affect Scale (PANAS) (Watson, Clark, & Tellegen, 1988).

Need to Transform Raw Data. As simple as the study sounds, analysis is not easy. All puzzles were solved. The main performance score was the time that a dyad took to do each puzzle. But raw puzzle time is influenced by three factors that are not relevant to the goal of the study which was to ascertain how performance varies as a function of a dyad's history of interaction. Since the game is played in English, English proficiency can affect performance as a whole. Game-playing skill can also influence performance as a whole. Since role-play gaming is complex, training does not necessarily make all players equally proficient. Furthermore, individual puzzles are not equally difficult (see the boxplots in Figure 6), so simply comparing performance across different puzzles does not fairly indicate the temporal evolution of relative performance. Hence, the raw times needed to be transformed prior to analysis.

Data Preparation: Covariates. In a similar fashion as Warren(2008), dyad English proficiency and game experience were treated as covariates and the effects were removed using multiple regression.

Data Preparation: Unequal Difficulty of Puzzles. After removal of the effects of English proficiency and gaming-experience, the adjusted times still have a distribution similar to Figure 6. Although the puzzles are known prior to testing, in contradistinction to the demographic covariates, it is still virtually impossible to devise a set of complex engaging cooperation-dependent puzzles of equal difficulty. This is not a weakness or defect but rather reflects in virtual reality a hard fact of real life: Real world social-interaction cooperative tasks are inherently unequal in difficulty.

Re-scoring performance on unequal tasks in order to enable comparisons must therefore make use of statistical adjustment methods. Within each puzzle type, raw solution times were normalized to have a mean of 0 and a standard deviation of 1. Since high numbers meant slower times, the normalized scores were then inverted about zero so that the mean remained 0, the standard deviation remained 1, but high numbers now meant better performance in the sense that quicker solutions are better. Finally, the new scores were transformed to have a mean of 50 and a standard deviation of 10. These Relative Adjusted Quickness (RAQ) scores (Warren, 2011) can then be compared across puzzles.

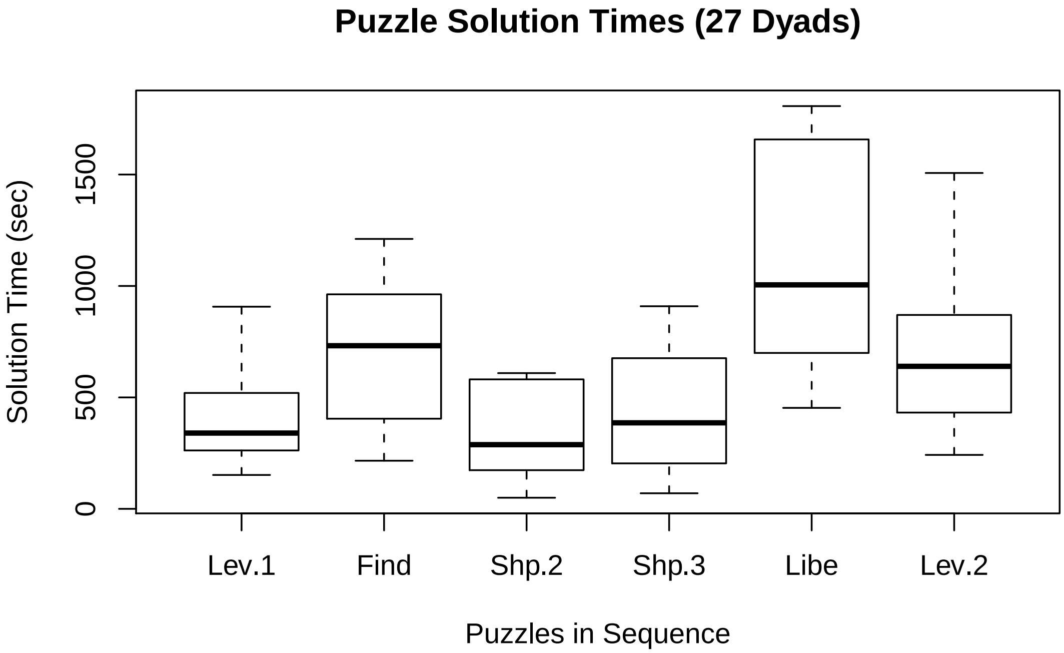

Figure 6: Raw puzzle completion times.

Data Preparation: Determining Trend Curves. A sequential plot of the RAQ scores per dyad shows considerable variability. Since what matters with respect to the hypothesis illustrated in Figure 5 is the trend, quadratic trend curves were computed for each of 27 dyads. For three types of dyad cultural mix, Figure 7 shows the mean quadratic trend lines of the RAQ scores in the temporal sequence in which the puzzles were administered. The three types of dyad cultural mixes were: Same-culture dyads, mixed-culture dyads, and dyads from a single nation but at least one of whose members self-identified as being from a sub-culture within the home nation.

Discussion. None of the curves in Figure 7 resembles the notional curves in Figure 5.

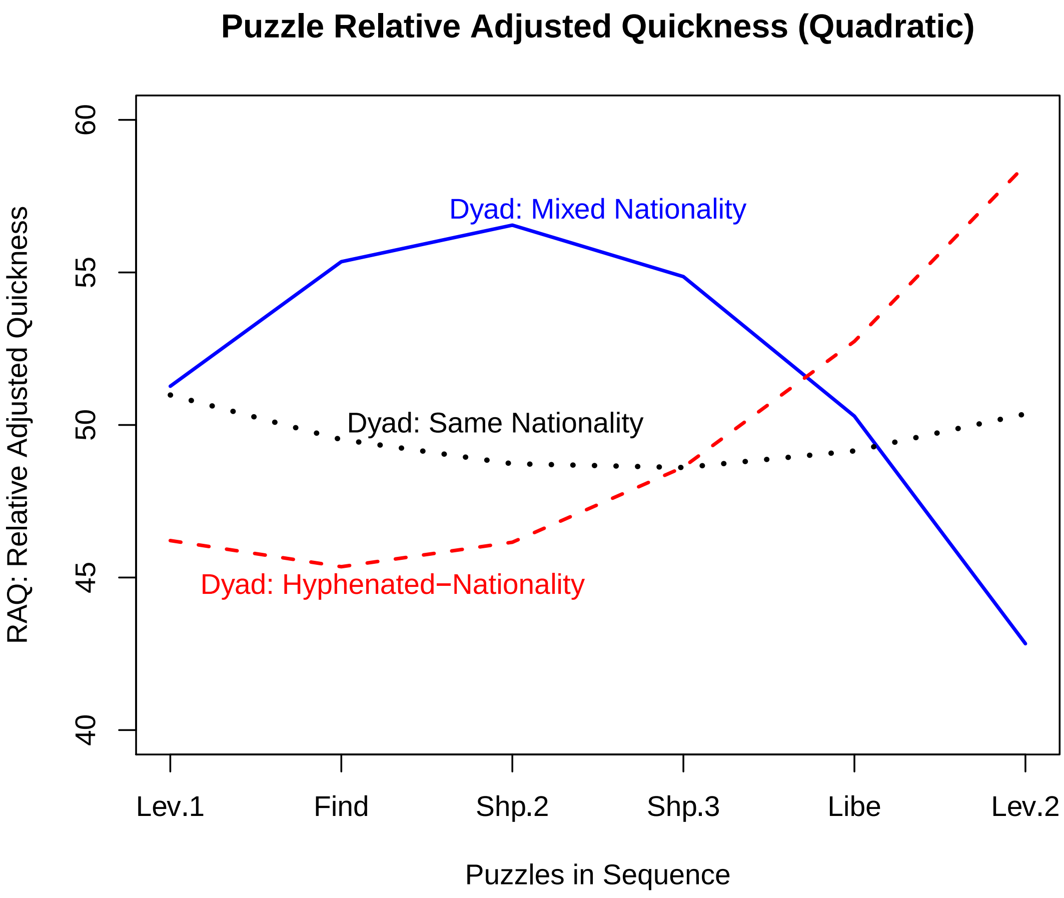

Figure 7: Puzzle Relative Adjusted Quickness

For the mixed-nationality dyads, the progressive deterioration in solution performance from the third puzzle onwards was not unexpected. But the initial rise in performance up to the third puzzle is puzzling. Perhaps there is an initial positive attitude and it takes a while for the difficulties in communicating and interacting to begin to have a negative effect.

The slightly bowl-shaped performance curve for the same-nationality dyads was also not unexpected. In retrospect, it can be argued that a third curve, flat and unchanging, should be added to Figure 5: When communication is matter-of-fact and straight forward in a dyad, there is no reason to expect any significant change in performance. Some such dyads might be better than others, but whatever their "baseline" level of performance no change over time would be expected. The realization that some dyads will differ on their baseline levels of performance suggests another change to the hypothesis embodied in Figure 5, namely that the vertical axis should reflect *relative change* in sequential performance not the absolute level. There is no reason whatsoever to think that all teams would start at the mean level of performance.

The generally increasing performance of the partially-mixed dyads was also contrary to expectations. Perhaps positive effects of diversity are operating.

At end of the period of performance covered in this report, additional data was being collected and more extensive analyses were pending. Hence the current results and discussion are not complete or final. the current analyses, however, have suggested several ideas to be pursued in the future analyses. Some of these question the premises of the current study. There has been an underlying assumption that the greater the cultural difference in a dyad, the poorer the communication, and hence, a growing deterioration in effectiveness over time when dealing with complex tasks.

Cultural differences do affect team performance, but which differences matter and to what degree are not well understood. Considerable distance on many cultural dimensions can exist between strong friends and allies. Also, some of the worst conflicts occur in civil wars and internecine fighting in which the combatants are culturally and otherwise almost identical.

Furthermore, team performance is not static but depends on the cumulative experiences of the team as it goes about solving problems. Contrary to the tone of the arguments leading to Figure 5, the history of experiences of a team and the time-history of its effectiveness are not inevitably monotonically increasing or decreasing, but rather may have considerable "ups and downs." The ups-and-downs are evident in Figure 7 in spite of the use of quadratic regression to smooth them out.

A major contribution of this study was the development of techniques for a suitable data analysis. The results indicate that mixed-culture dyads did exhibit improved performance over time whereas the same-culture dyads tended not to show any great change over time (Warren, 2011). These conclusions are somewhat simplified, however, as the actual picture is more complex. The research is continuing beyond that reported for the current report.

6 MODELING CULTURAL EFFECTS

In addition to statistical analyses of the data collected in CCL, the research program has included several modeling efforts to better understand the results of the experiments and culture in general.

6.1 Emergence of Complex Cultural Beliefs

Although complex culture beliefs are relatively stable, they are not immutable. As people have experience with their environment and exchange information with other people about their experiences, their beliefs can be modified and change over time. Upal and Warren (2009) used an agent-based simulation to model the emergence of complex cultural beliefs. An interesting aspect of the simulation is that agents, when they met along pathways, could query each other about what the environment was like where they had come from. Since the agents were in competition to accumulate resources, it was sometimes advantageous for an agent to lie. An experiment is planned to compare the performance of humans with that of the computer agents.

6.2 One-Factor Model of Cultural Effects

The rationale of the NATO-HFM-138 study was that good planning, resource allocation, and situational awareness depend on good communication and coordination. In turn, good communication and coordination are facilitated by sharing a common culture. Hence, the principal hypothesis was: Homogeneous-culture teams (i.e., teams whose members are all from the same nation) perform better than mixed-culture teams (i.e., teams whose members are from different nations). To aid analysis, Warren (2010) made the tacit logic chain that culture is efficacious solely via the single mechanism of communication explicit in the model shown in Figure 8. As previously noted, the data did not support this model.

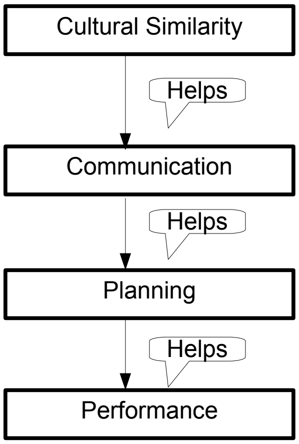

Figure 8: Implicit single-factor model of cultural effects.

6.3 Two-Opposing Factors Model of Cultural Effects

In addition to asking what the effects of cultural diversity on team-performance are, we can also ask about how—and when—these effects come about. In particular, after considering possible theoretical explanations for the superior performance of culturally-mixed teams, Warren (2010) speculated that the diversity of the mixed-culture teams permitted better planning and produced a better search strategy in the sense of the *Wisdom of Crowds* (Surowieki, 2004/2005).

The Factors by Which Diversity Might Operate. This presumptive facilitating effect of diversity on search strategy clashes with the presumptive facilitating effect of homogeneity on communication and coordination. Moreover, whatever the relative "strengths" of the opposing factors, the possible mechanisms by which the putative factors might operate are not symmetric:

- The effects of diversity are arguably due to the interaction of the team *as a whole* in a planning phase, whereas
- The consequences of the quality of communications arise from the many individual one-on-one conversations which take place in the post-planning execution phase.

Under this analysis, culture has its effects via two mechanisms. Further, these mechanisms are presumed to be in opposition: As one promotes, the other hinders success (and *vice versa*). And still further, these mechanisms are presumed to operate in different phases. Two example scenarios help illustrate the extremes:

- A well-coordinated team might attempt to execute a poor plan.
- A team which develops a great plan might bungle its execution.

Figure 9 illustrates this two-factor opposing-mechanism model of how cultural-diversity affects team performance.

Domain of Applicability & Complexity Considerations. The domain of applicability of the model thus assumes that the team's task spans two separate phases, namely, a

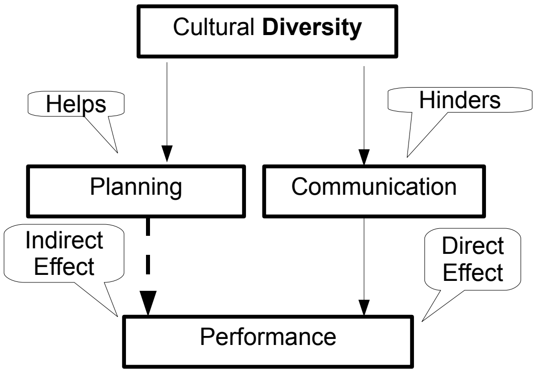

Figure 9: Two-factor opposing-mechanism model of cultural-diversity effects on a complex team task.

planning phase and an execution phase. But a task that requires planning prior to execution suggests that a certain high-level of complexity is involved. Complexity here does not refer to the number of interacting agents, but rather to the nature of the interactions and their consequences. These interactions can be quite different in the various phases:

- Training: A task complex enough to require extensive training raises the question of training effectiveness. When homogeneous teams come from diverse cultures or when a team itself is culturally diverse, there is no guarantee that all teams are equivalently trained even when the same procedures and criteria are used.

- Planning: The task must be complex enough that planning is not just a matter of choosing between a few well-defined procedural alternatives. In the NATO experiment, teams had to determine role assignments and responsibilities, allocate resources, determine how to conduct the search, and formulate policies for dealing with unplanned events.

- Execution: The task facing the team must be complex enough that it cannot be carried out by independent agents. Success must require communication, coordination, and the asking for and giving of assistance. In the complex urgent real-world situations to which this model is intended to apply, it is possible for calls for help to not be heard, or for a potential help-provider to not be able to immediately respond or even not to be able to respond at all.

- Task objectives: An appropriate task to which the model applies need not have clear, well-defined objectives. Since many complex real-worlds tasks are not well-defined, defining a realistic objective can be considered a pre-planning phase with which a team must concern itself. In fact, it can be argued that the more fuzzy or ill-defined the objectives, the greater the potential benefit from having a culturally-diverse team.

General Considerations & Comments. A few points should be clarified or emphasized:

- Diversity has its *effect* in the execution phase (since *all* effects are made manifest in the execution phase), but it is presumed due to the discussions in its planning phase and the “policies” and procedures that the team adopts.

- It is explicitly assumed that any communication problems of culturally diverse teams have a negligible impact during the planning phase. It is the (presumed) greater richness of ideas put forth and considered that matters, and it is assumed that the participants share enough of a common language for the ideas to “come across” no matter how awkwardly phrased, heavily “accented,” or haltingly expressed from the viewpoint of native-speakers of the common language.

- However, awkward phrases, heavy accents, and hesitations while someone searches for a word can have negative impacts during the execution phase. These types of miscommunications cover anything from mis-hearings to misunderstandings.

Testing the Two-Factor Model. The need to test the two-factor model has led to two very different lines of investigation. In one, Liu and Warren (2009) used fuzzy decision trees to explore cultural influences on the team planning and communication which lie at the heart of the model. The second arose within Warren (2010) while designing an agent-based simulation to test the model: How to build teams of agents such that the teams has specific levels of cultural diversity? Discussion of thee two explorations follow.

6.4 Team Planning and Communication: Fuzzy Decision Trees

To further understand both the NATO data and the assertions of the two-opposing-factors model of the effects of cultural diversity, Liu and Warren (2009) used a completely different type of modeling approach than agent-based simulation. Instead, they used fuzzy decision trees and information visualization techniques to study the effects of cultural diversity on team planning and communication—but not on team task performance. The reason for focusing on planning and communication is that the two-factor opposing-mechanism model of cultural effects proposed by Warren (2010) differentiates between a planning stage (in which fluid communication is not critical) and an active execution stage (in which facile communication is critical). For this analysis, diversity was not based on the national composition of the four-member teams, but rather on each individual member's responses on a questionnaire measuring the Hofstede (2001) cultural dimensions. Team cultural diversity was indexed using the population standard deviation of the four team-members's scores on a particular cultural dimension.

The results show that team planning quality was best predicted by diversity in the cultural dimension of power distance followed by diversity in the cultural dimension of masculinity. Team communication quality was best predicted by diversity in masculinity and then by diversity in power distance. A significant contribution of this model is that it teases out which particular aspects of cultural diversity are important and when. That is, cultural diversity is not a univariate dimension, but rather exists or not along many different cultural dimensions.

6.5 Diversity Considerations: Team Diversity Index

A model of the effects of team cultural diversity requires an index of its central concept.

Culture is a set of characteristics shared, more or less, by a group of people. Many individual dimensions have been proposed as particularly salient and useful for contrasting and comparing groups. These include power distance (e.g., Hofstede, 2001), Analytic-Holistic thinking (e.g., Nisbett, 2003) and Individualism-Collectivism (e.g., Triandis, 1995). It is sometimes useful to further differentiate some dimensions such as Institutional Collectivism versus In-group Collectivism (e.g., House, Hanges, Javidan, Dorfman, & Gupta, 2004).

For these, and many other, proposed cultural dimensions, researchers have developed various questionnaires, sets of rating scales, and other tests. These measuring "instruments" are usually given to *individuals* and scored first to determine an individual's position along the cultural dimension in question. The distribution of one group's scores can then be compared with the distribution of another group's scores.

What is important for assessing team diversity is that scores on some dimension, or multidimensional composite, are available for each individual member of a team. Assume, then, that we have individual scores on some cultural dimension for all members of a team. How should we combine the scores to form a team diversity index?

Suitable Team Diversity Indexes. Suitable team cultural diversity indexes include the standard deviation and the mean absolute difference. Surprisingly, Warren (2009) discovered that the Gini coefficient, often used to measure inequality in economics, and the relative mean difference are not suitable diversity indexes since they can yield different values depending the direction of the skewness of scores even though two teams may be identical in diversity.

The simplest team diversity indexes would be to use the (sample) standard deviation (SD) of the individual culture-index values or the mean absolute difference (MAD). If the values were all equal, the SD's or MAD's would be zero indicating no cultural diversity—no matter what the team size—as expected.

For reasons given in Warren (2009), a better metric is to use the fraction (or percent) that a team index is of the *maximal possible team index* given the particular range of values and the team size:

$$\text{team diversity index} = \frac{\text{actual team index}}{\text{maximum possible team index}} \tag{1}$$

Advantages of percent of maximal possible index-value as as team diversity index include:

- Avoiding the problem of the dependency of a simple index value on the scale or range of the measuring instrument and the dependency on team size.
- Statements like “10% diverse” and “80% diverse” are intuitively grasped and permit ready comparisons.

There are, unfortunately, some problems with using Equation 1. Warren (2009) also discovered that procedures for determining maximal value of the standard deviation and mean absolute deviation are not in the literature. He did determine that the maximal values for the standard deviation and the mean absolute difference are one-half of the scale range for teams with even numbers of members. If the number of members is odd, there is an adjustment given in Warren (2009).

Forming Teams with Maximal Diversity. Warren (2009) further discovered that the team composition that has the maximum diversity as indexed by the standard deviation or the mean absolute difference requires half the members to score at one extreme and the other half to score at the other extreme if the team size is an even number. If the team size is odd, the remaining member must be in the middle for maximum standard deviation but may be anywhere if using the mean absolute difference.

Parenthetically, this procedure, although technically correct, does not necessarily have complete intuitive appeal. A person might expect that a team with culture scores of [0 33 67 100] would be maximally diverse rather than a team with scores of [0 0 100 100]. That is, a team with scores spread evenly across its range can be argued to be more diverse that a team with half its scores at one extreme and half at the other extreme. The search for other—and more intuitive—indexes of diversity is continuing.

Use of a Diversity Index in Modeling and Simulation. The reason for developing a measure of maximal diversity and a procedure for determining a team composition having maximal diversity was to be able to conduct rigorous simulation test of models of cultural effects on interaction and performance.

When testing with humans, the values of diversity are only slightly under an experimenter's control. An experimenter might pair people from the same culture or from different cultures in the hope that the resulting culture value are similar or different, but there is no guarantee of success. It is unlikely that the extremes will be represented, and hence the data will always be vulnerable to technical restriction-of-range effects.

With agent-based simulation, an experimenter can create agents with any desired value on a cultural dimension. This enables the use of powerful experimental designs in which the full range of possible values may be explored such as in Warren (2009) and (2010).

7 DISCUSSION & CONCLUSIONS

The empirical and modeling research in the Culture & Cognition Laboratory has focused on the effects of differing cultural compositions on the effectiveness of teams engaged in cognitively-demanding tasks requiring extensive communication and cooperation for success. In contrast to most cultural research which is conducted with paper-and-pencil measures (Brouwers et al., 2004), the CCL research program has featured action-demanding interactive activity using an immersive role-play computer game. Since research using serious games is still relatively new and scant, there are numerous lessons to be learned with respect to conducting such experiments (Bainbridge, 2010; Sherwin, 2007). In addition to lessons about the effects of culture on cognition, these technical lessons are worth noting since the validity of the theoretical contributions depends on the quality of the data from which the conclusions are drawn and the models designed to explain.

The NATO RTO HFM-138/RTG and the dyad experiments are conceptually simple but very complex with respect to methodological aspects such as the role-play game itself, details of the scenario, and a teams task and options. The experiment was also complex logistically both within a session and throughout the entire experiment. Numerous lessons have been learned within the broad categories of conception, the game itself, methodology, logistics & execution, and analysis.

7.1 Concepts, Hypotheses, & Theoretical Issues

The experiments used a complex computer game to study the effects of different cultural compositions in multinational coalitions and cooperation in dyads. This is appropriate due to the inherent and pronounced immersive quality of such games, but also due to the fact that tomorrow's military recruits are growing up playing more and more such games and developing computer and communications skills not typical of people from a generation ago. Questions about what make some teams more effective than others are difficult to answer in general, but differential computer experience adds a fresh and urgent dimension to these questions about team adaptability especially in multinational coalitions and geographically-distributed teams..

7.2 The Game & Its Characteristics

As discussed earlier, the game used in CCL is based on a complex, very absorbing, and immersive role-play game, Neverwinter Nights. Both the general SABRE research environment and the specific scenarios were extensively piloted and iteratively refined, and numerous lessons learned concerning the games and games in general apply in the development phase and the execution phase of the research.

- Features that permit creativity, variant behaviors: The game and task were chosen so as to permit a large degree of creativity and self-determination by the teams in how they would approach accomplishing their mission. But the more degrees of freedom given the teams and the more unstructured the task, the less control that the experimenters have and the harder it is to interpret the various results. It should be noted that the experimental scenario was relative "static" in that there were no surprises or major incidents occurring during the game-play. The use of non-briefed events could certainly be introduced into the game, but we chose not to do so to maintain a degree of comparability in the experiment.

- Main tasks versus side quests: Although the scenario was relatively "static" as just discussed, their were some opportunities for teams to engage in "side quests" (such as helping a non-player character computer-generated girl search for a lost pet) which could garner goodwill points but which would take time away from the main task. Such side-quests do add realism and permit opportunities for non-routine decision making.

- Experimenter's viewpoint and used and unused game features: From the experimenter's viewpoint, the game can be very rich in decision making opportunities. However, a particular team might decline various opportunities or not be very creative and thus, as the game unfolds, the game can evolve into something less rich because certain avenues are not explored.

- Player's viewpoint: By observing the players and from their comments after the experiment, it is clear that the game succeeded in being immersive and absorbing. Players did not report trying to figure-out what the experiment was about, but rather quickly became fully engaged in the task at hand.

- Experimenter interaction/intervention possibilities: The underlying game (Neverwinter Nights) has a dungeon-master mode feature in which a game-master (or an experimenter) can have an invisible "avatar" (i.e., personal representative character in the game) which can interact with the game environment and other characters. This feature was needed only on the rare occasions when a human players avatar got "stuck" in a wall (there are occasional glitches since the software is very complex) to free the avatar without the humans awareness. One lesson learned is to be prepared for such events and to know how to deal with them.

- A related lesson for future research is that the dungeon-master mode can be utilized to introduce some player-action-contingent events into the game-play. For example, a door could be closed (by the unseen dungeon master) thereby trapping the human player until they radio for rescue by another player. Such in-game or in-line modifications require active monitoring and in-game intervention by an experimenter, but the possibilities are intriguing.

- Underlying & unused game features: Since the underlying game permits many behaviors which are not needed or allowed in a particular scenario (i.e., casting spells), it is important to prevent their accidental use by, for example, disabling the right-mouse button.

7.3 Methodology

The game and scenario used were complex to learn and complex to play, but the permitted behaviors are manifold. This richness means that certain methodological aspects that are normally under an experimenters complete control in a more traditional laboratory experiment are not-controlled or even non-controllable. Some methodological lessons learned or special problems encountered in conducting the study are:

7.3.1 Subjects

- Incomparability of subject pools: When subjects come from multiple countries, it is very difficult to be sure that the subject pools are comparable. For example, a junior officer in one country might be considered a student in a second country and hence not in the pool of the second country.

- Size of subject pool: In spite of the size of many militaries, the pool of available military subjects can be surprisingly small. Military officers, in particular, are busy people and often have critical jobs from which they can not be spared for a block of four to six hours. When constraints are placed on the characteristics of an entire team, such as requiring a certain age range, the effective size of the pool can, and does, shrink drastically. This also applies to studies using foreign students at a university since the numbers of willing subjects is small.

- Representativeness of subjects to intended application: Military officers have specialized occupations and some of these are not interchangeable. A medical officer cannot be expected to perform the work of a pilot. When the pool of possible subjects is small, allowance must be made to permit more people to qualify for the experiment. Unfortunately, this means that the relevance of the results to the target population could become compromised.

- Team formation: Within a country and within the same research site, some individuals might know each other and some might be strangers. But teams whose members have a common past history can be expected to function differently than teams whose members are strangers. A background question about prior knowledge of or experience with other team members should be included along with the demographic questions.

- Distributed "team" issues & considerations: When team members come from different geographic locations or even nations, there are special issues of team formation and identification with the team. This problem is compounded when the only interaction team members can have is via a keyboard. But, however cumbersome "introductions" and interactions might be among distributed teams, such teams are becoming more and more common in real-world operations.

7.3.2 Non-Player Characters

Role-play games are not a traditional experimental vehicle. Hence, they involve new categories for complete methodological description. Computer-generated non-player characters are like human subjects in some ways and quite unlike humans in ways that can be exploited to gain more information from the human players at a reasonable cost.

- Non-player characters (NATO study): The town populace was comprised of computer-generated "non-player characters" (NPC's). The avatars of the human players could interact with the NPC's via scripted question and answer sets. The NPC's were programmed to make a variety of responses such as providing tips regarding the whereabouts of suspicious activity. But some NPC's could lie (i.e., they were programmed to provide false information). NPC's have great potential in general for research purposes. This area needs more work, but one which can bring rich rewards especially as the NPC's take on theoretically-based or empirically-grounded personality and cultural characteristics. The number, content, and veracity of messages should be addressed by any researcher.

- Non-player characters (Dyad study): Non-player characters play a role even in the dyad experiment. In addition to their use in training the human subjects, a key role is played by an NPC who guards a critical door and will not allow passage until a specific item is offered. Use of such devices helps keep the humans searching. Another advantage of NPC's is that they help populate the virtual world so that it is not empty of other "people." This adds a bit pf "realism" to the environment but in a controlled and inexpensive fashion.

7.3.3 Experimental Design

- As discussed above, the pool of potential subjects can be very small. Thus, it is imperative to use as efficient an experimental design as possible with respect to the number of necessary subjects.

- The experiment must also be very efficient with respect to its time demands. Six hours makes it hard to get subjects and also can be a strain on the subjects. The total amount of time includes time for pre- and post-game questionnaires. These need to be kept to a minimum.

- Statistical design, matched samples, controlled & uncontrolled variables: Another consequence of the limited subject pool is that there are few possibilities for matching subjects on extraneous variables or for assigning subjects to pre-specified levels in a factorial design on factors such as age. Indeed, Warren (2008) has argued that full experimenter control over all variables of interest in a complex experiment is not just difficult but actually impossible. However, this does not mean that the effects of the confounding variables such as computer-game experience, English proficiency, or other covariates cannot be assessed. Using analysis of covariance (ANCOVA) and other regression-based techniques, these effects can be measured and then be statistically partialled-out.

7.3.4 Procedures

- Questionnaires: The use of a computer game does not obviate the use of more traditional 5- or 7-point rating scales. Both both pre- and post-game questionnaires are needed for obtaining such information as demographic data and personality and cultural profiles.

- In-game probes: Another feature that recommends use of the game is the occurrence of in-game probes. For example, in the NATO study on three occasions, a "superior officer" (wholly within the game), probed the subjects with questions relating to their situation awareness. The use of in-game probes can be a powerful tool and is a supplement to the out-of-game questionnaires and the in-game situations (which are themselves tests).

- Training: different learning curves and times: There were two training phases, one in which individuals learned basic one-person actions such as moving forward, picking up objects, using a map, using one's journal, etc., and a second phase in which an individual learned to communicate with others. People were permitted to complete basic individual-action training at their own pace. But this meant that people finished basic training at different times. Fast learners often had to wait a long while at an in-game waiting area while slower learners were still mastering basic skills. The in-game waiting area had amusing activities to keep people busy, but it could be a long time, and the amusement nature of the filler activities could contribute to a sense that the overall game was not a serious exercise.

- Training: proficiency criteria and removal concerns: Related to the problem of different people taking time to reach a sufficient level of proficiency is the question at what level to set the proficiency criteria. Although this never occurred in the main experiments, there was a case during piloting with non-military subjects when one individual simply could not achieve sufficient skill to enable that experimental run to continue. Since this occurred during piloting, no time limit had been set, and this led to a boredom problem with the other three players. Of course, not only do such aborted sessions waste peoples' time, it can be costly in terms of money since (non-military) subjects still have to be paid.

- Local testing issues: breaks etc. When testing was at one site in the NATO study, the procedure was to conduct pre-questionnaire completion, individual, and team testing phases before lunch. The planning and search phases were after lunch, but this raises the chances that some forgetting might take place. It is now recommended that a short refresher training session occur after lunch. Eliminating the need for a major break was one reason for the development of the dyad study.

- Distributed testing: time zones consideration: The mixed-nation testing was done over the Internet. But since the experiment spanned 6 time zones and could take 6 hours, the experiment began relatively early in the morning for the Americans and ended relatively late at night for the Europeans. The previous point's reference to lunch has to be modified, but the issue of the timing of breaks becomes even more important. Anything that lengthens the experiment, such as the above recommendation for a refresher training phase, must be carefully weighed against the effect of a long day on some peoples performance.

7.4 Administration & Logistics

- Subject scheduling issues: As discussed earlier, the size of the pool of possible subjects was severely limited. One administrative difficulty that resulted from this was that of being able to schedule at least four people for a test day. It often took considerable effort on the part of the research team to locate and enlist the minimum of four people needed for a team. This was another reason for the development of the dyad scenario.

- The difficulties were great enough that there were times when a session had to be canceled in advance due to either the inability to locate four subjects or due to the advance cancellation by one of the volunteers. This again put a burden on the research team to contact the remaining volunteers.

- No-shows: Even on days when four people had been scheduled, there was the all too common and exacerbating problem of a scheduled volunteer not appearing and thus forcing the cancellation of the session and the attendant loss of time of those who did appear for the experiment.

  - One technique for dealing with the problem of no-shows is to schedule more people than required. Due to the limited subject pool, this option was difficult to exercise.
  - Even if there was a large pool to routinely overbook subjects, overbooking does not guarantee that the required number of subjects will show up. The reality of research on teams is that no-shows are all too common: If six people are scheduled for a four-person session, only three might show up.
  - But overbooking has is own problems. One problem is that if all show up for the experiment, some method has to be used to determine whom to dismiss and in such a way that the excused person is treated with respect and made to feel that their effort is still appreciated and not wasted.
  - In research without the need for subjects with highly specialized characteristics, one way to not waste any “unusable” subjects who report for an experiment (either too few or too many) is to have alternate lower-priority experiments ready which can use whatever number of people are available after due consideration for the needs of the highest priority experiment. However, this was not an option due to the small size of the pool. Any potential subjects who could not be run even after they reported for the experiment needed to be asked to reschedule if at all possible.

- Scheduling a long experiment over an ocean: The mixed team portion of the experiment often required having an American and Bulgarian on the same team. Arranging for a short meeting across five or more time zones is hard, but arranging an experimental session that will take six or more hours means that Europeans will be finishing quite late in their day and that Americans will be starting quite early in their day. The definition of lunch break is thus relative and has to be taken into account when the potential subjects are given details about what is being asked of them when they are solicited.

- Computer operators and local administrators: The above remarks about long experiments across an ocean also apply to the local computer operators and local experiment administrators who, by the nature of their responsibilities, must be present both before and after the subject session.

- Internet operators: The mixed team portion of the NATO experiment also required the use of a knowledgeable team of SABRE experts and an internet operations center to host and coordinate the multi-site internet portion of the experiment. In order to ensure smooth operations and prevent loss of precious data, the internet operations had to be flawless. This required much advance preparation and testing of communication links and procedures. Although given scant mention in the experimental write-up and methods sections of the reports, this aspect of the experiment is crucial and required considerable effort.

7.5 Data Collection, Processing & Analysis

The SABRE testbed features automatic data collection of both pre-game questionnaires and within-game activity and communications. SABRE also collates the data from the various individual team sessions and collates the data into large spreadsheet files for post-processing by various statistical packages.

- Although SABRE does provide some basic statistics, it was felt best to leave the main analyses to the various members of the experimental teams and the statistical packages they prefer. One reason for this is the large and diverse nature of the data recorded and the subsequent opportunities for post-experiment data mining. The datasets resulting from this experiment will yield rich treasures as they continue to be mined.

- With a data set resulting from the game-play and questionnaires of 224 subjects, it is inevitable that there will be some missing data. Since different analysts have different preferences for dealing with missing data, it is imperative that there be tight configuration management of the raw and early-processed data sets that are distributed to the various analysts. In turn, it is also important that the various analysts maintain their own processed-data file configuration management with full description of the decisions they made and the procedures they followed.

7.6 Drawing Conclusions & Making Recommendation

In spite of running 224 subjects in the NATO study, the resulting number of four-person teams was 56.

- Since the analyses are all team-centric, the conclusions are based on the relatively small number of 56 teams. As such, statistical power is weak and the conclusions must be taken with caution.

- Also, as discussed by Warren (2008), there are several confounds that also serve to temper conclusions and recommendation such subject differences in age, computer-game experience, and English proficiency.

- However, the confounds are, to a large degree, unavoidable due to the complex nature of the subject populations. They are not deficiencies in the experimental design. Fortunately, there are statistical techniques such as ANCOVA and linear regression which can partial-out the effects of the confounds and enable the drawing of confound-free conclusions.

7.7 Final Comments

We have partial answers to what effects cultural differences have on team performance in cognitively demanding situations entailing cooperation. But, in general, much of what makes a team cooperative and successful in a multinational coalition is still not fully understood. However, the value of using an immersive computer game to provide rich data sets to help provide such answers has been amply demonstrated. As tomorrows military recruits become more and more experienced with complex immersive computer games than the recruits of yesterday, it becomes imperative that we study the possible impact of such experience on selection and training for tomorrows more computer-reliant military.

To reiterate a key point about cultural distance made earlier, cultural differences do affect team performance, but which differences matter and to what degree are not well understood. Considerable distance on many cultural dimensions can exist between strong friends and allies. Also, some of the worst conflicts occur in civil wars and internecine fighting in which the combatants are culturally and otherwise almost identical.

Another key lesson made earlier is that team performance is not static but depends on the cumulative experiences of the team as it goes about solving problems. Contrary to the tone of the arguments leading to Figure 5, the history of the experiences of a team and the time-history of its effectiveness are not inevitably monotonically increasing or decreasing, but rather may have considerable "ups and downs." The ups-and-downs are evident in Figure 7 in spite of the use of quadratic regression to smooth them out.

Lastly, the data sets collected under this workunit are rich and complex. The analysis, data mining, and behavior modeling will continue under other efforts.

REFERENCES

Bainbridge, W.S., (2010). *The Warcraft civilization: Social science in a virtual world.* Cambridge, MA: MIT Press.

Bioware Corp. (2004). *Neverwinter Nights*^{TM} - Platinum Edition (Computer Game Software). New York: Atari Interactive.

Brouwers, S.A., Van Hemert, D.A., Breugelman, S.M., & Van de Vijer, F.J.R. (2004). A historical analysis of empirical studies published in the *Journal of Cross-Cultural Psychology*, 1970–2004, *Journal of Cross-Cultural Psychology, 35*, 251–262.

Choi, I., Koo, M., & Choi, J.A. (2007). Individual differences in analytic versus holistic thinking. *Personality and Social Psychology Bulletin, 33,* 691–705.

Fischer, R. (2004). Standardization to account for cross-cultural response bias. *Journal of Cross-Cultural Psychology, 35*, 263–282.

Hofstede, G. (2001). *Cultures consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). Thousand Oaks, CA: Sage.

House, R.J., Hanges, P.J., Javidan, M., Dorfman, P.W., & Gupta, V.] (Eds.). (2004). *Culture, leadership and organizations: The GLOBE study of 62 societies.* Thousand Oaks, CA: Sage.

Liu, Y., & Warren, R. (2009). Using fuzzy decision trees and information visualization to study the effects of cultural diversity on team planning and communication. In Dana Nau and Aaron Mannes, (Eds.), *ICCCD 2009: Proceedings of the Third International Conference on Computational Cultural Dynamics* (pp. 45–53). Menlo Park, CA: Association for the Advancement of Artificial Intelligence. ISBN 978–1–57735–443–7.

Leung, A., Diller, D., & Ferguson, W. (2005). SABRE: A game-based testbed for studying team behavior. Proceedings of the Fall Simulation Interoperability Workshop (SISO). Orlando, FL, September 18-23, 2005. Available from www.sisostds.org as paper 05F-SIW-047.

NATO RTO HFM-138/RTG (2008). Report of the NATO Research and Technology Organization (RTO) Human Factors and Medicine Panel Research Task Group 138 on "Adaptability in Multinational Coalitions. Brussels: NATO.

Nisbett, R. E. (2003). *The geography of thought: How Asians and Westerners think differently . . . and why.* New York: The Free Press.

Sherwin, J. (2007). Get a [Second] Life studying behavior in a virtual world. *Observer, 20,* June/July, pp. 18–21. (A publication of the Association for Psychological Science)

Surowiecki, J. (2004/2005). *The wisdom of crowds.* New York: Anchor Books.

Triandis, H.C. (1995). *Individualism and collectivism.* Boulder, CO: Westview Press.

Upal, M.A., Warren, R. (2009). Simulating the emergence of complex cultural beliefs. In T. Terano, H. Kita, S. Takahashi, & H. Deguchi (Eds.), *Agent-based approaches in economic and social complex systems V: Post-proceedings of the AESCS International Workshop 2007*, (17–28). Tokyo: Springer.

Warren, R., Sutton, J., Diller, D., Ferguson, W., & Leung, A. (2004). A game-based testbed for culture & personality research. Proceedings of the NATO Modeling and Simulation Group - 037 Workshop: Exploiting Commercial Games for Military Use. The Hague, The Netherlands, 20-21 Oct 2004.

Warren, R., Diller, D.E., Leung, A., Ferguson, W., & Sutton, J.L. (2005). Simulating scenarios for research on culture & cognition using a commercial role-play game. In M.E. Kohl, N.M Steiger, F.B. Armstrong, and J.A. Jones, (Eds.), *Proceedings of the 2005 Winter Simulation Conference*. Orlando, FL.

Warren, R. (2008). Mixed- & homogeneous culture military team performance on a simulated mission: Effects of age, game-experience, & English proficiency. Proceedings of the NATO RTO HFM–142 Symposium on Adaptability in Coalition Teamwork, held in Copenhagen, Denmark. Brussels: NATO.

Warren, R., & Sutton, J. (2008). Using a computer game for research on culture and team adaptability: Lessons learned from a NATO experiment. In the *Proceedings of the NATO Research and Technology Organization (RTO) Human Factors and Medicine Panel HFM–142 Symposium on Adaptability in Coalition Teamwork*. Copenhagen, Denmark (April 2008).

Warren, R., (2009). Designing maximally, or otherwise, diverse teams: Group-diversity Indexes for testing computational models of cultural and other social-group dynamics. In Dana Nau and Aaron Mannes, (Eds.), *ICCCD 2009: Proceedings of the Third International Conference on Computational Cultural Dynamics* (pp. 78–85). Menlo Park, CA: Association for the Advancement of Artificial Intelligence. ISBN 978–1–57735–443–7.

Warren, R., (2010). Modeling direct and indirect effects of cultural diversity. *International Journal of Computational Intelligence: Theory and Practice, 5*, 37–48.

Warren, R., (2011) Evolving problem-solving performance by mixed-culture dyadic teams. In P. Vink & J. Kantola (Eds.), *Advances in occupational, social, and organizational ergonomics*, (289–298). Boca Raton, FL: CRC Press.

Watson, D., Clark, L.A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54,* 1063–1070.

LIST OF ACRONYMS

AFRL	Air Force Research Laboratory
AFRL/HE	AFRL Human Effectiveness Directorate
ANCOVA	Analysis of Covariance
BBN	BBN Technologies, Inc.
CCL	Culture & Cognition Laboratory
CRDA	Cooperative Research and Development Agreement
DMSO	Defense Modeling and Simulation Office
MAD	Mean Absolute Difference
NATO	North Atlantic Treaty Organization
NATO RTO	NATO Research and Technology Organization
NATO RTO HFM	NATO RTO Human Factors and Medicine Panel
NPC	Non-Player Character
SABRE	Situation Authorable Research Environment
SD	Standard Deviation
WSU	Wright State University